Tècniques de Feature Weighting per casos no supervisats. Implementació a GESCONDA

M. Sànchez-Marrè[a], S. Gómez[b], F. Teixidó[c], K. Gibert[d]

[a] Knowledge Engineering and Machine Learning group, UPC.
[b] Dep. Arquitectura de Computadors, UPC.
[c] Dep. Informàtica. Enginyeria i Arquitectura La Salle, URL.
[d] Dep. Estadística i Investigació Operativa, UPC

Juliol 2006

Índex

Capítol 1

Feature Weighting

1.1 Aspectes teórics del Feature Weighting

El *Feature Weighting*[**?**] és una evolució natural del Feature Selection, tècnica que determina quin conjunt d'atributs classifica millor un conjunt d'instàncies, descartant així un subconjunt dels atributs.

Feature Weighting es fa servir per augmentar la rellevància dels atributs. L'objectiu principal d'aquesta tècnica és assignar pesos més elevats als atributs més rellevants i pesos més baixos als atributs irrellevants o que no ens aporten un valor.

Davant d'un exemple és normal pensar que hi ha atributs més rellevants que altres, es a dir, que hi haurà atributs que són característics d'un conjunt d'exemples i altres atributs que són característics d'un altre conjunt. Aquests atributs ens interessa que prenguin major importància o tinguin més pes a l'hora d'avaluar la similitud entre els diferents exemples.

El procés de *Feature weighting* és aplicable a molts paradigmes diferents. En aquesta memòria només estudiem el comportament de tres algorismes concrets.

Amb el *Feature weighting* fem un pas més, ja que l'atribut ja no és actiu o inactiu sinó que se li assigna un grau de significació definit dins l'interval de $[0-1]$, sent 0 el valor de menor significació i 1 el de major significació. Cal observar, doncs, que qualsevol tècnica de *Feature Selection* és un subconjunt del *Feature Weighing* on els pesos assignats als atributs deseleccionats equival a 0 i als atributs seleccionats val 1.

La rellevància, en el camp de l'aprenentatge automàtic, pot definir-se de diferents formes, però una definició tradicionalment acceptada és la següent:

Rellevància X_i és rellevant si i només si existeix un x_i i y tal que

$$P(X_i = x_i) > 0$$
$$\text{i}$$
$$P(Y = y | X_i = x_i) \neq P(Y = y)$$

És a dir, X_i és rellevant si conèixer el seu valor pot canviar la estimació de Y o, en altres paraules, si Y és condicionalment dependent de X_i.

Existeixen altres definicions més precises que distingeixen entre rellevància forta o feble o d'altres que fan referència a la rellevància de variables, però no necessitem entrar tant en detall.

1.2 Feature Weighting a Dominis Supervistats

La característica principal dels algorismes de determinació automàtica de pesos es que fan servir l'etiqueta de classe per a fer l'assignació de pesos. L'objectiu és assignar pesos alts als atributs altament correlacionats amb l'atribut de classe.

En aquest camp podem distingir dos tipus d'algorismes:

Mètodes d'assignació de pesos global Assumeixen que la rellevància d'un atribut és la mateixa per a tot l'espai d'instàncies o per a tots els valors de l'atribut.

Mètodes d'assignació de pesos local No realitzen la suposició que fan els mètodes globals i assignen pesos locals a diferents regions dels valors d'un atribut.

1.3 Feature Weighting a Dominis No Supervisats

La investigació de Feature Weighting en dominis no supervisats és escassa, en canvi, per dominis supervisats hi ha molts mètodes diferents que ho implementen. La comunitat té la creença que en dominis no supervisats els resultats havent aplicat qualsevol tècnica de *Feature weighting* no poden millorar.

Una primera aproximació per a dominis no supervisats seria fer un clustering de les instàncies, assignar etiquetes artificials i llavors aplicar mètodes de Feature Weighting per a dominis supervisats. Clarament, aquest mètode està esbiaixat pel procés de clustering.

Una altre aproximació, en absència de l'etiqueta de classe, la distinció de rellevància dels atributs és derivar segons diferents criteris, com ara:

- Correlació.
- Variança.
- Entropia.

A continuació, mostrem diferents algorismes que s'han implementat per la realització de la memòria de *Feature Weighting* per a dominis no supervisats.

1.3.1 Gradient-Descent Technique (GD)

Aquesta proposta consisteix principalment a definir un index d'avaluació d'atribut E[1]. Aquest mètode utilitza el càlcul dels gradients per obtenir els màxims i mínims del conjunt d'exemples i així poder discernir quins són els atributs més rellevants. La finalitat és arribar a un index E prou baix i apte. Cal dir que mentre la similitud entre diferents instancies tendeix a 1 o 0, l'index E tendeix a 0.

L'Algorisme **??** mostra el pseudo-codi corresponent a aquest mètode.

while *E(w) < llindar* **do**
  forall *Attribute* A_i **do**
    forall *Attribute* A_k **do**
      SM^1_{ik} := SM(A_i,A_k, 1) ;
      SM^w_{ik} := SM(A_i,A_k, w) ;
      E(W) := estimació (SM^1_{ik}, SM^w_{ik}) ;
    end
    $\Delta W := -\eta \frac{\partial \mathrm{E}}{\partial w_f}$
  end
end

Algorithm 1: Algorisme GD.

Cal dir que:

- SM^1_{1k} és la similitud entre dues instàncies sense aplicar cap pes
- SM^w_{1k} és la similitud entre dues instàncies aplicant els pesos associats als atributs en aquell moment.

i el valor de ΔW es calcula aplicant les formules **??**, **??** i **??**, on:

- χ^2 és la distancia entre els valors de l'atribut j
- *alpha* i η prenen valors de l'interval [0..1]
- N és el nombre d'instàncies

$$\frac{\partial \mathrm{E}(w)}{\partial w_f} = \frac{2\left[\sum_{p=1}^{N}\sum_{q=1,q\neq p}^{N}(p - 2SM^1_{pq}\frac{\partial SM^w_{pq}}{\partial d^w{}^{pq}}\frac{\partial d^w_{pq}}{\partial w_f}\right]}{N(N-1)} \quad (1.1)$$

$$\frac{\partial SM^w_{pq}}{\partial d^w_{pq}} = \frac{-\alpha}{\left(1+\alpha d^w_{pq}\right)^2} \quad (1.2)$$

$$\frac{\partial d^w_{pq}}{\partial w_f} = \frac{w_j\chi^2_j}{\sqrt{\sum_{j=1}^{N} w_j^2\chi_j^2}} \quad (1.3)$$

[1] veure definició d'aquest índex al paper *Ïnstance-Based Learning Techniques of Unsupervised Feature Weighting do not perform so badly!"*

1.3.2 Unsupervised Entropy Based-1 (UEB-1)

Aquest mètode es basa en el càlcul d'entropies. L'algorisme calcula totes les entropies del conjunt de dades eliminant en cada iteració un atribut diferent, és a dir, a la primera iteració desactivarà el primer atribut i calcularà l'entropia del conjunt, a la segona iteració desactivarà el segon atribut i calcularà la nova entropia i així successivament fins arribar a l'calcular totes les entropies desactivant un atribut cada vegada. Per K atributs es calcularan K entropies.

Una vegada es tenen les entropies que genera l'eliminació d'un atribut del conjunt, es pondera l'atribut segons l'entropia que genera normalitzada. El cas extrem seria un atribut que no genera entropia és un atribut que se li assignarà un pes de 0, en canvi, un atribut que genera una entropia màxima, se li assignarà un pes de valor la unitat. Recordem que els pesos estan normalitzats, prenen valors del rang [0..1].

L'Algorisme **??** mostra el pseudo-codi corresponent a aquest mètode.

Aquest algorisme és molt senzill d'entendre si prenem la definició d'entropia.

Definició 1 (Entropia) *Quantitat de desinformació*

forall *Atribut* A_i **do**
| $e[i] :=$ EntropiaEliminantUnAtribut(A_i) ;
end
forall *Atribut* A_i **do**
| $W_i := \frac{e[i]-\min(e)}{\max(e)-\min(e)}$;
end

Algorithm 2: Algorisme UEB-1.

1.3.3 Unsupervised Entropy Based-2 (UEB-2)

Aquest segon mètode, el qual també es basa en el càlcul d'entropies, aplica una política de recompensa.

Un pes inicial de 0.5 es assignat a cada atribut i es calcula la entropia per a tota la base de dades. Llavors, de la mateixa forma que en el UEB-1, es calcula una nova entropia després d'haver esborrat un dels atributs cada vegada. Aleshores, si el valor de la nova entropia (havent esborrat un atribut) és menor que la entropia de tota la base de dades, es penalitza l'atribut disminuint en 0.1 el seu pes; si el valor de la nova entropia es major, llavors es recompensa l'atribut augmentant el seu pes en 0.1.

Aquest cicle es repeteix fins que arriben a una condició de convergència o, més senzill, N vegades.

L'Algorisme **??** mostra el pseudo-codi corresponent a aquest mètode. L'algorisme no ho mostra, però per evitar pesos negatius o més grans que 1 en cas de moltes iteracions, cal fer una normalització dels pesos obtinguts en el rang [0..1].

```
forall Atribut A_i do
 |  W_i := 0.5 ;
end
e_o := Entropia();
while ¬ condició do
 |  e_n := 0;
 |  forall Atribut A_i do
 |  |  e_n = EntropiaEliminantUnAtribut(A_i);
 |  |  if e_n < e_o then
 |  |  |  W_i := W_i − 0.1;
 |  |  else
 |  |  |  W_i := W_i + 0.1;
 |  |  end
 |  end
end
```

Algorithm 3: Algorisme UEB-2.

1.4 Petit Exemple: *Iris*

Hem realitzat un exemple pràctic amb una base de dades *Iris* de 150 instàncies eliminant l'atribut de classe.

Els resultats els podem veure a la Taula **??**, on els *pesos* $\in [0..10]$. Si comparem els resultats podem veure com la tendència és assignar major rellevància a l'amplada i a la llarga del pètal. En el cas de l'algorisme UEB-2 les assignacions de pesos sempre resulten més extremes (valors 0.0 o 10.0).

Atribut	**GD**	**UEB-1**	**UEB-2**
Amplada del pètal	10.0	7.5155	10.0
Amplada de sèpal	10.0	0.0	0.0
Llargada de sèpal	10.0	2.4608	0.0
Llarga del pètal	10.0	10.0	10.0

Taula 1.1: Assignació automàtica de pesos amb els algorismes GD ($\alpha = 0.1$, $\mu = 0.2$ i $threshold = 1 \times 10^{-5}$), UEB-1 i UEB-2 (5 iteracions) als atributs de la base de dades *Iris*.

Capítol 2

Objectius

2.1 Propósits

L'objectiu principal d'aquest projecte és incorporar al programari *Rule Induction* de tres tècniques o mètodes que implementen el *Feature Weighting*. Aquestes tècniques són interessants des del punt de vista que assignen pesos utilitzant diferents criteris a un conjunt de dades d'entrada.

En aquest cas, però, el propòsit és incorporar i aplicar aquests tres mètodes a les dades provinents d'una depuradora d'aigües residuals i estudiar la seva dels atributs.

Cal recordar que una vegada descoberts els pesos pels diferents mètodes, i gràcies al programari *Rule Induction* podríem aplicar diferents mètodes de *clustering* o analitzar les dades sempre tenint en compte els atributs més importants. En definitiva *Feature Weighting* és un bon mecanisme per eliminar soroll del conjunt de dades d'entrada.

Els tres mètodes implementats i agregats al programari *Rule Induction* són els següents:

1. GD
2. UEB-1
3. UEB-2

Un cop disposem dels algorismes de *Feature Weighting* i haguem executat diferents testos per validar la seva correctesa estarem en disposició de:

1. Executar els tres algorismes amb dades provinents de la depuradora i determinar els pesos dels atributs.
2. Fer una anàlisi descrivint les característiques dels resultats generats per cadascú dels algorismes.

3. Confirmar amb el coneixement d'un expert que els pesos assignats per cadascú dels algorismes als atributs del conjunt de dades són correctes, o almenys, "a ull", ho semblen.
   Com que ens serà impossible o molt difícil realitzar aquest pas, i ja que disposem d'una descripció de les dades relativament precisa (veure Capítol **??**), intentarem nosaltres mateixos interpretar la rellevància de les dades i confirmar que els algorismes de Feature Weighting implementats generen els resultats esperats.

4. Realitzar diferents processos de clustering amb les dades de la depuradora fent servir els pesos calculats pels nostres algorismes de clustering i sense pesos, analitzant llavors les diferencies en les classes generades, si és que n'hi ha. Seria també interessant analitzar si hi ha moviment de classe d'algunes instàncies degut a la incorporació dels pesos.
   En aquest cas, no disposar del software de clustering necessari per a realitzar aquesta tasca ens deixarà aquest punt com una tasca a realitzar en un futur.

2.2 Requeriments secundaris

A més a més de la completa integració dels algorismes de Feature Weighting en el software *Rule Induction*, el desenvolupament d'aquests algorismes tenen certs requeriments secundaris.

Adjuntem en aquest grup de requeriments aquells que són necessaris pel correcte desenvolupament del programari o bé aquells que són requeriments propis d'interfície demanats per l'usuari. A continuació hi donem una llista:

1. Utilitzant el codi com a gairebé única font de documentació [1], comprendre el model de dades que hauríem de fer servir en els nostres algorismes, la gestió de la interfície gràfica per a la correcta integració de les nostres finestres, etc.

2. Generació dels menus corresponents per a indicar l'execució d'un procés de Feature Weighting i de les finestres tant d'interacció amb l'usuari i de captació de paràmetres com de presentació dels resultats.

3. Permetre la exportació dels pesos calculats per un algorisme a fitxers de text pla. En concret, a fitxers en format `"*.csv"`, que a la pràctica són fàcilment exportables a fulles de càlcul.

4. Permetre la importació de pesos des de fitxers de text pla (el format corresponent a l'exportació `"*.csv"`). El càlcul del Feature Weighting és un procés amb un elevat cost temporal, així que permetre la importació estalviarà haver de repetir els càlculs si aquests es guarden en fitxers.

[1] Codi que inicialment no compilava en la seva totalitat (problema que hem resolt nosaltres) i que no disposava del codi font d'una de les seves classes.

2.3 Aplicacions

L'aplicació directa del Feature Weighting és determinar la rellevància dels atributs, per fer-ho, com ja s'ha dit anteriorment, s'usa un pes associat a cadascun dels atributs del conjunt d'exemples d'entrada. Però també es pot fer servir com a eina de Feature Selection. Els algorismes afegits permeten determinar per un conjunt de dades no supervisades, en quina mesura són de rellevants els atributs del conjunt. D'aquesta manera es podria desactiva amb el programari *Rule Induction* aquells atributs on els algorismes de Feature Weighting hagin determinat que tenen un pes per sota d'un llindar definit de forma arbitrària per l'usuari.

Capítol 3

Presentació de les Dades

3.1 Introducció

La descripció de la planta per a cada día consisteix a caracteritzar el caudal d'entrada, l'estat de la mescla després del primer decantador, el caudal de sortida i l'estat de la mescla al reactor biológic. En aquesta caracterització es fan servir principalment mesures de volum i resultats d'anàlisis químics i biológics.

A continuació es detallen les variables que s'han fet servir a la descripció de les dades.

3.2 Variables d'entrada

Q-E Caudal d'entrada. Es mesura en metres cúbics d'aigua per dia ($\frac{m^3}{dia}$). El valor mínim observat a les dades és de 20500 i el màxim de 54088.6.

FE-E Pretactament amb ferro. Es mesura en mil.ligrams de ferro per litre d'aigua ($\frac{mg}{l}$). El valor mínim observat a les dades és de 89.8 i el màxim de 8.

PH-E Mesura del pH a la entrada. El pH es una mesura de la activitat dels ions d'hidrogen en una solució i, per tant, mesura la seva acidesa. El pH només pren valors entre 0 i 14, peró a les dades observades el valor mínim és de 7.2 i el mínim de 8.

SS-E Sòlids en suspensió a l'entrada. Es mesura en mil.ligrams de sòlids per litre d'aigua ($\frac{mg}{l}$). El valor mínim observat a les dades és de 62 i el màxim de 655.

SSV-E Sòlids volàtils en suspensió a l'entrada. Es mesura en mil.ligrams de sòlids per litre d'aigua ($\frac{mg}{l}$). El valor mínim observat a les dades és de 19 i el màxim de 593.

DQO-E Fracció de materia orgànica degradable per acció d'agents químics oxidants sota condicions d'acidesa a l'entrada. Es mesura en mil.ligrams d'oxígen per litre d'aigua ($\frac{mg}{l}$). El valor mínim observat a les dades és de 27 i el màxim de 1579.

DBO-E Fracció de materia orgànica biodegradable en aigua residual a l'entrada. Es mesura en mil.ligrams d'oxigen per litre d'aigua ($\frac{mg}{l}$). El valor mínim observat a les dades és de 69 i el màxim de 987.

NKT-E Suma del amoni i el nitrogen orgànic a l'entrada. Es mesura en mil.ligrams d'oxigen per litre d'aigua ($\frac{mg}{l}$). El valor mínim observat a les dades és de 17.9 i el màxim de 82.

NH4-E Amoni a l'entrada. Es mesura en mil.ligrams d'nitrogen per litre d'aigua ($\frac{mg}{l}$). El valor mínim observat a les dades és de 8.1 i el màxim de 347.

P-E Fósfor a l'entrada. Es mesura en mil.ligrams de fósfor per litre d'aigua ($\frac{mg}{l}$). El valor mínim observat a les dades és de 2.1 i el màxim de 24.

DBO/DQO-E Quocient de materia orgànica biodegradable en aigua residual a l'entrada. No té unitats de mesura. El valor mínim observat a les dades és de 0.15 i el màxim de 1.04.

3.3 Variables després de la decantació

PH-D Mesura del pH a la decantació. El pH es una mesura de la activitat dels ions d'hidrogen en una solució i, per tant, mesura la seva acidesa. El pH només pren valors entre 0 i 14, peró a les dades observades el valor mínim és de 7.1 i el mínim de 7.9.

SS-D Sòlids en suspensió a la decantació. Es mesura en mil.ligrams de sòlids per litre d'aigua ($\frac{mg}{l}$). El valor mínim observat a les dades és de 40 i el màxim de 192.

SSV-D Sòlids volàtils en suspensió a la decantació. Es mesura en mil.ligrams de sòlids per litre d'aigua ($\frac{mg}{l}$). El valor mínim observat a les dades és de 13 i el màxim de 134.

DQO-D Fracció de materia orgànica degradable per acció d'agents químics oxidants sota condicions d'acidesa a la decantació. Es mesura en mil.ligrams d'oxigen per litre d'aigua ($\frac{mg}{l}$). El valor mínim observat a les dades és de 27 i el màxim de 538.

DBO-D Fracció de materia orgànica biodegradable en aigua residual a la decantació. Es mesura en mil.ligrams d'oxigen per litre d'aigua ($\frac{mg}{l}$). El valor mínim observat a les dades és de 36 i el màxim de 274.

NKT-D Suma del amoni i el nitrogen orgànic a la decantació. Es mesura en mil.ligrams d'oxigen per litre d'aigua ($\frac{mg}{l}$). El valor mínim observat a les dades és de 15.1 i el màxim de 74.

NH4-D Amoni a la decantació. Es mesura en mil.ligrams d'nitrogen per litre d'aigua ($\frac{mg}{l}$). El valor mínim observat a les dades és de 8.9 i el màxim de 37.1.

DBO/DQO-D Quocient de materia orgànica biodegradable en aigua residual a la decantació. No té unitats de mesura. El valor mínim observat a les dades és de 0.15 i el màxim de 1.

Com a primer anàlisi, si comparem els valors màxims i mínims de les variable d'entrada amb les de la decantació, podem observar com els valors, en general, decreixen, denotant així la neteja que ja s'ha realitzat a l'aigua.

3.4 Variables de sortida

PH-S Mesura del pH a la sortida. El pH es una mesura de la activitat dels ions d'hidrogen en una solució i, per tant, mesura la seva acidesa. El pH només pren valors entre 0 i 14, però a les dades observades el valor mínim és de 7 i el mínim de 8.

SS-S Sòlids en suspensió a la sortida. Es mesura en mil.ligrams de sòlids per litre d'aigua ($\frac{mg}{l}$). El valor mínim observat a les dades és de 2.8 i el màxim de 174.8.

SSV-S Sòlids volàtils en suspensió a la sortida. Es mesura en mil.ligrams de sòlids per litre d'aigua ($\frac{mg}{l}$). El valor mínim observat a les dades és de 1.6 i el màxim de 134.8.

DQO-S Fracció de materia orgànica biodegradable en aigua residual a la sortida. Es mesura en mil.ligrams d'oxigen per litre d'aigua ($\frac{mg}{l}$). El valor mínim observat a les dades és de 9 i el màxim de 163.

DBO-S Fracció de materia orgànica biodegradable en aigua residual a la sortida. Es mesura en mil.ligrams d'oxigen per litre d'aigua ($\frac{mg}{l}$). El valor mínim observat a les dades és de 2 i el màxim de 84.

NKT-S Suma del amoni i el nitrogen orgànic a la sortida. Es mesura en mil.ligrams d'oxigen per litre d'aigua ($\frac{mg}{l}$). El valor mínim observat a les dades és de 2 i el màxim de 67.

NH4-S Amoni a la decantació. Es mesura en mil.ligrams d'nitrogen per litre d'aigua ($\frac{mg}{l}$). El valor mínim observat a les dades és de 0.5 i el màxim de 31.5.

P-S Fósfor a la sortida. Es mesura en mil.ligrams de fósfor per litre d'aigua ($\frac{mg}{l}$). El valor mínim observat a les dades és de 0.6 i el màxim de 7.

DBO/DQO-S Quocient de materia orgànica biodegradable en aigua residual a la sortida. No té unitats de mesura. El valor mínim observat a les dades és de 0.07 i el màxim de 1.

Novament, podem veure el decreixement general dels valors de les variables a la sortida, és a dir, l'aigua cada cop està, en general, més neta.

3.5 Variables del tractament biològic

V30-B Anàlisis volumètric 30; qualitat de sedimentació de la mescla. Es mesura en mil.lilitre per litre d'aigua ($\frac{ml}{l}$). El valor mínim observat a les dades és de 77 i el màxim de 770.

MLSS-B Sòlids en suspensió del licor mescla. Es mesura en mil.ligrams de sòlids per litre de licor mescla ($\frac{mg}{l}$). El valor mínim observat a les dades és de 754 i el màxim de 3294.

MLVSS-B Sòlids volàtils en suspensió del licor mescla. Es mesura en mil.ligrams de sòlids per litre de licor mescla ($\frac{mg}{l}$). El valor mínim observat a les dades és de 185 i el màxim de 2100.

IM-B Quocient entre V30 i MLSS-B. Es mesura en mil.lilitre per gram ($\frac{ml}{g}$). El valor mínim observat a les dades és de 58.5 i el màxim de 577.

CM1-B Fracció de materia orgànica degradable (kg) per acció d'agents químics oxidants per sòlids en suspensió (kg). El valor mínim observat a les dades és de 0.02 i el màxim de 1.43.

CM2-B Fracció de materia orgànica biodegradable (kg) per sòlids en suspensió (kg). El valor mínim observat a les dades és de 0.05 i el màxim de 3.86.

MCRT-B Edat cel·lular. Es mesura en dies ($dies$). El valor mínim observat a les dades és de 1.78 i el màxim de 341.99.

QB-B Caudal del reactor biològic. Es mesura en metres cúbics d'aigua per dia ($\frac{m^3}{dia}$). El valor mínim observat a les dades és de 19883 i el màxim de 52244.6.

3.6 Altres variables

QR-G Caudal de recirculació. Es mesura en metres cúbics d'aigua per dia ($\frac{m^3}{dia}$). El valor mínim observat a les dades és de 17932.6 i el màxim de 49527.

QP-G Caudal de la purga. Es mesura en metres cúbics d'aigua per dia ($\frac{m^3}{dia}$). El valor mínim observat a les dades és de 0 i el màxim de 1080.

QA-G Afluència d'aire. Es mesura en metres cúbics d'aigua per dia ($\frac{m^3}{dia}$). El valor mínim observat a les dades és de 96451 i el màxim de 367840.

TRH-C Temps de resistència hidràulica. Es mesura en hores (h). El valor mínim observat a les dades és de 2.16 i el màxim de 9.63.

3.7 Aspectes generals

Un aspecte important en la descripció de les dades és la quantitat de valors nuls que contenen. En el nostre model de dades, cal destacar que hi ha un conjunt de variables que presenten una quantitat de valors nuls considerable. Aquestes variables són:

- NKT-E, NKT-D i NKT-S.
- NH4-E, NH4-D i NKT-D.
- P-E, P-D i P-D.

Un altre aspecte curiós és el de les variables QP-G (cabal de la purga) i FE-E (pretactament amb ferro). Aquestes variables presenten en diferents moments un seguit de mesures amb valor 0 que s'allunyen de la tendència general. Això podria ser degut a un mal funcionament temporal de l'aparell que realitza la mesura o que durant un temps ha calgut tancar el cabal de purga o aturar el pretactament amb ferro.

3.8 Anàlisi Gràfic

A la Figura **??** (esquerra) podem veure una comparativa dels boxplots que generen les tres variables que fan referència als sòlids en suspensió. Podem veure com a l'entrada (SS-E) es troben els valors més alts, com decreixen en arribar a la decantació (SS-D) i com arriben a uns valors mínims a la sortida (SS-S).

El mateix efecte el podem observar amb d'altres variables, com és el cas de la variable SSV, a la Figura **??** (dreta), la variable pH, a la Figura **??** (esquerra) o la variable DQO, a la Figura **??** (dreta).

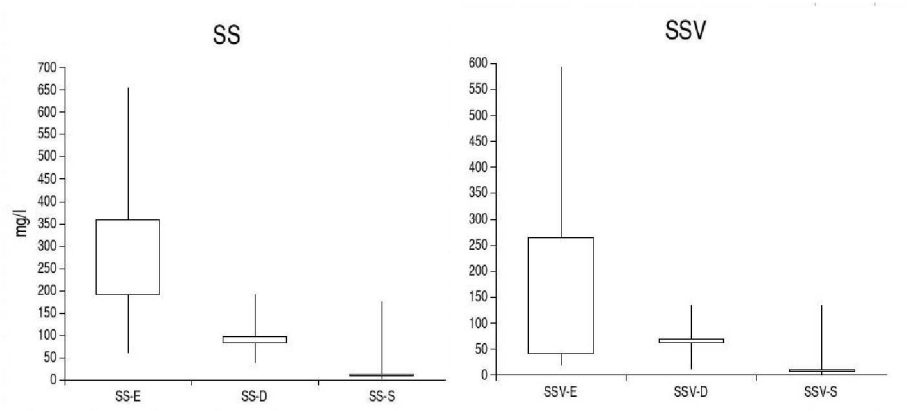

Figura 3.1: Boxplot de les variables SS-{E,D,S} i SSV-{E,D,S}.

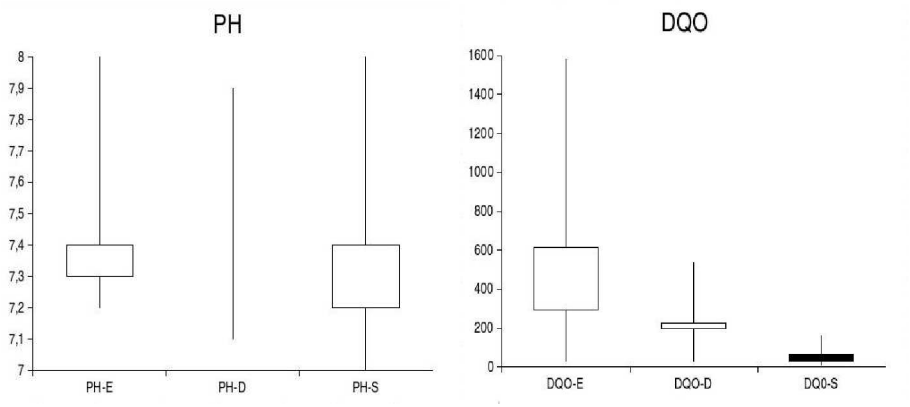

Figura 3.2: Boxplot de les variables PH-{E,D,S} i DQO-{E,D,S}.

Capítol 4

Desenvolupament

4.1 Tasques de desenvolupament

Per poder fer l'experimentació s'ha tingut que afegir noves classes que incorporessin tot el programari desitjat i funcionés en sintonia amb el programari ja implementat. És a dir, s'ha implementat el següent:

- Diàlegs per la interacció amb l'usuari
- Diàlegs de presentació resultats
- Diàlegs de visualització
- Algorismes de les tècniques anteriorment exposades
- Afegir funcionalitats no contemplades al programari

4.2 Diàlegs

S'han desenvolupat els següents diàlegs:

- *CVDialegFeatureWeighting*
- *CVDialegPresentarPesosAtributs*
- *CVDialegFixaFuncioDistancia*

4.2.1 CVDialegFeatureWeighting

El diàleg *CVDialegFeatureWeighting* Ens mostra tots els algorismes de *Feature weighting* implementades, i ens dona la possibilitat d'ajustar els paràmetres necessaris.

Els algorismes que ens ofereix són:

- *UEB-1*

Figura 4.1: Captura del diàleg per la selecció del mètode de *feature weighting*

- *UEB-2*
- *GD*

Per l'algorisme *UEB-1* no permet ajustar cap paràmetre degut a que aquest mètode no se li pot ajustar ningun.

L'algorisme *UEB-2* se li poden ajustar les iteracions que ha de fer l'algorisme. És l'únic paràmetre que rep.

Per l'últim mètode tenim la possibilitat d'ajustar diversos paràmetres: la η, el factor d'aprenentatge α i el llindar de tolerància, en el diàleg expressat en forma de mantissa i exponent, degut al valors que pren. Veure [?] [?] pels detalls dels valors que poden prendre.

4.2.2 CVDialegPresentarPesosAtributs

Aquest diàleg ens mostra els resultats obtinguts de l'execució del mètode de *Feature Weighting* executat. El resultat es mostre en dues columnes: La primera columna indica el nom de l'atribut i la segona columna ens mostra el pes assignat per l'algorisme seleccionat.

A la figura **??** es pot veure una captura del diàleg *CFDialegPresentarPesosAtributs*. El diagrama de classes corresponent a aquest diàleg es troba a la figura **??**.

Amb els pesos resultants d'una execució podem fer tres operacions, tal i com es pot observar a la figura **??**. Sempre podrem:

Figura 4.2: Diagrama de classes del diàleg *CVDialegFeatureWeighting*

Figura 4.3: Captura diàleg de presentació de resultats, *CVDialegPresentarPesosAtributs*

- Acceptar, els pesos s'assignaran als atributs corresponents per a futurs càlculs
- Rebutjar, els pesos queden totalment descartats
- Exportar, els pesos calculats es poden exportar a un fitxer de text per a futures utilitzacions d'aquests

rebutjar i exportar a fitxer

4.2.3 CVDialegFixaFuncioDistancia

Per tots els mètodes de *Feature weighting*, també, és important fixar la funció de distancia a utilitzar. Aquest diàleg contempla les següents:

- Euclidea
- Manhattan
- Clark
- Canberra
- L'eixample, podent assignar el paràmetre α corresponent
- Minkowski, podent assignar el paràmetre r corresponent

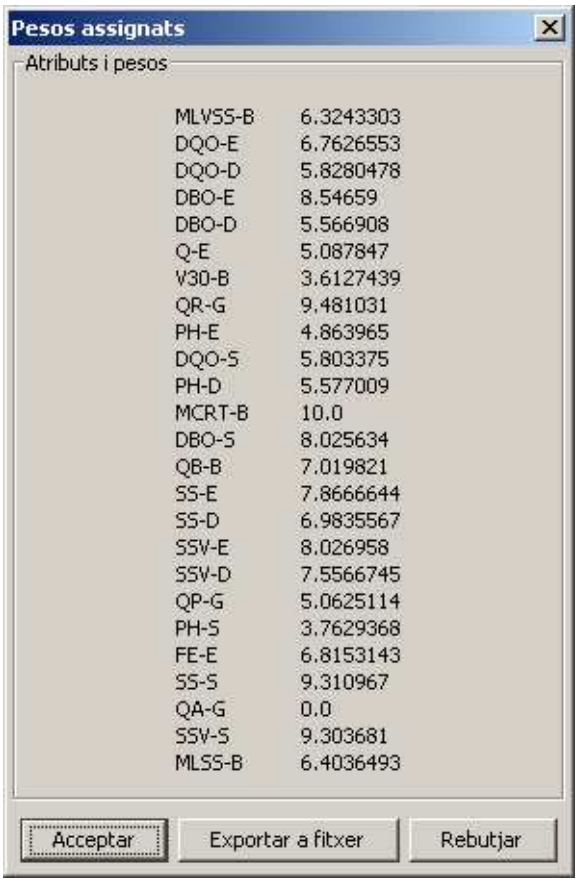

Figura 4.4: Diagrama de classes pel diàleg, *CVDialegPresentarPesosAtributs*

Figura 4.5: Captura diàleg per fixar la funció pel càlcul de distàncies, *CVDialegFixaFuncioDistancia*

- Cosinus

La figura **??** ens il·lustra el diàleg *CVDialegFixaFuncioDistancia*. El diagrama de classes es pot veure a la figura **??**.

4.3 Accions

Aquestes són classes implementades per suportar i dirigir el pes dels events. En aquest cas s'han afegit al programari dos tipus d'accions:

- Per capturar els events i executar els mètodes de *Feature weighting* implementats
- Per fixar la funció distància a executar

El diagrama de classes corresponent a les classes implementades es pot veure a la figura **??**.

Figura 4.6: Diagrama de classes del diagrama per fixar la funció de càlcul de distàncies, *CVDialegFixaFuncioDistancia*

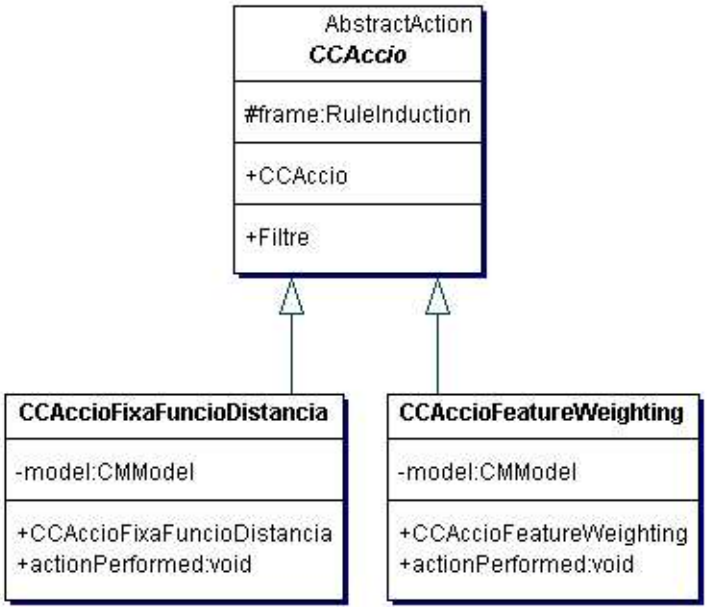

Figura 4.7: Diagrama de classes de les accions

4.4 Model

En el paquet model és on s'han implementat totes les funcionalitats demanades, aquestes són:

- Mètode *UEB-1*
- Mètode *UEB-2*
- Mètode *GD*

A part s'ha afegit un mètode a la classe *CMModel*. La funció afegida és *setDistancia*, la qual és útil per fixar una distància al model. P.e. tots els càlculs de l'algorisme del gradient *GD* es basen en la distància Euclidea.

Aquests algorismes estan implementats en només una classe, aquesta es pot veure a la figura **??** i la classe *CMModel* final queda il·lustrada a la figura **??**.

Els mètodes implmentats estan bastament explicats en els punts **??**, **??** i **??**.

Figura 4.8: Classe *Model* final

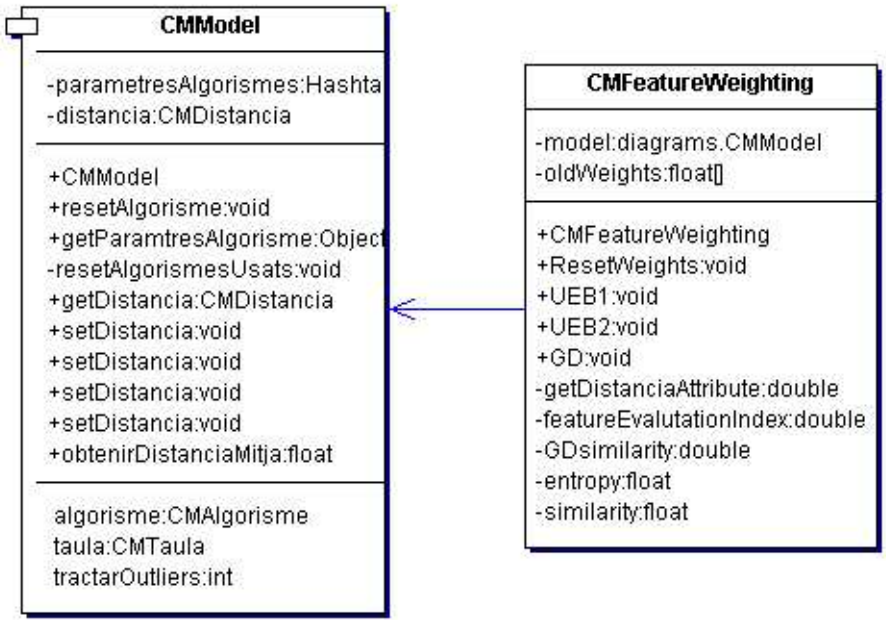

Figura 4.9: Diagrama de classes *Feature weighting*

Capítol 5

Experimentació

5.1 Descripció dels Experiments

En aquest capítol mostrem els resultats obtinguts amb l'execució dels tres algorismes implementats al software *Rule Induction*: GD, UEB-1 i UEB-2, amb diferents paràmetres si escau.

La base de dades utilitzada prové d'una depuradora i conté 396 instàncies, cadascuna amb 25 atributs. La descripció de les dades i els seus atributs està feta la Capítol **??**. En totes les simulacions hem fet servir la distáncia Euclidea quan era necessari calcular la distància entre instancies[1].

La quantitat de dades i l'elevat cost dels algorismes fa que els temps d'execució siguin força elevats, tal i com veurem.

5.2 Resultats GD

Després d'executar el mètode *Gradient Descendant technique* diverses vegades amb diferents conjunts d'entrada, hem pogut comprobar que els resultats obtinguts no eren significatius. Normalment pondera per igual tots els atributs on els altres algorismes implementats donen resultats lleugerament diferents.

A part, el temps d'execució és molt més elevat. Això ens ha portat a no executar aquest algorisme amb les dades de la depuradora.

5.3 Resultats UEB

A la Taula **??** tenim els resultats de les execucions dels algorismes UEB-1 i UEB-2, aquest últim amb 5 i 10 iteracions.

[1] Només hem fet les proves amb aquesta la funció de distància Euclidea, peró hem afegit al software *Rule Induction* tot el necessari per a que l'usuari pugui seleccionar una distància i associar-la al model. D'aquesta forma l'usuari sel.lecciona per a tots els algorismes, no només els de Feature Weighting, quina funció distància i paràmetres s'han de fer servir durant els càlculs de distància.

El primer que veiem és que tots tres donen més rellevància al mateix conjunt d'atributs. Peró així com el UEB-1 fa una assignació de pesos més dispersa en el rang [0 - 10.0], el UEB-2 fa assignacions més extremes i sempre assigna pesos amb valor 0 o 10.0. A la pràctica, el UEB-2 es comporta com un métode de Feature Selection.

El fet que el UEB-2 es comporti d'aquesta forma es degut a que al finalitzar el procés, s'aplica una normalització dels pesos obtinguts a l'algorisme. A l'inici de la implementació de l'algorisme, tots els atributs comencen amb pes 0.5 i, a cada iteració, mitjançant un heurístic basat en entropies, es recompensa o es penalitza a cada atribut amb 0.1. Si el número de iteracions és elevat (més de 5), es podria donar el cas que obtinguessim pesos negatius. Per evitar això, es fa una normalització dels pesos obtinguts en el rang [0 - 10.0]. A més a més, s'afegeix el fet que, per les característiques de les dades, els atributs que a cada iteració són recompensats són sempre els mateixos, és a dir, hi ha un conjunt d'atributs que sempre són recompensats (els que acaben tenint pes 10.0) i la resta són sempre penalitzats (i acaben tenint pes 0.0).

És per aquest motiu que acabem d'explicar que els resultats pel UEB-2 amb 10 iteracions són els mateixos que amb 5 iteracions. I, encara que no es mostren els resultats, amb més iteracions (20, per exemple), passa el mateix.

Si ens fixem una mica més, podem veure que el UEB-2 amb 5 iteracions assigna pes 10.0 a tots els atributs on el UEB-1 els hi assigna un pes superior a 7.5.

Respecte als temps d'execució, podem comentar el següent:

- El UEB-1 va trigar de l'ordre de 2 minuts.
- El UEB-2, amb 5 iteracions, va trigar uns 12 minuts. És a dir, de l'ordre de 5 vegades el UEB-1, tal i com era d'esperar.
- El UEB-2, amb 10 iteracions, va trigar uns 25 minuts. És a dir, de l'ordre de 2 vegades el que triga l'execucio amb 5 iteracions.

Atributs	**UEB-1**	**UEB-2**	
		5 iteracions	10 iteracions
MLVSS-B	6.3243303	0.0	0.0
DQO-E	6.7626553	0.0	0.0
DQO-D	5.8280478	0.0	0.0
DBO-E	8.54659	10.0	10.0
DBO-D	5.566908	0.0	0.0
Q-E	5.087847	0.0	0.0
V30-B	3.6127439	0.0	0.0
QR-G	9.481031	10.0	10.0
PH-E	4.863965	0.0	0.0
DQO-S	5.803375	0.0	0.0
PH-D	5.577009	0.0	0.0
MCRT-B	10.0	10.0	10.0
DBO-S	8.025634	10.0	10.0
QB-B	7.019821	0.0	0.0
SS-E	7.8666644	10.0	10.0
SS-D	6.9835567	0.0	0.0
SSV-E	8.026958	10.0	10.0
SSV-D	7.5566745	10.0	10.0
QP-G	5.0625114	0.0	0.0
PH-S	3.7629368	0.0	0.0
FE-E	6.8153143	0.0	0.0
SS-S	9.310967	10.0	10.0
QA-G	0.0	0.0	0.0
SSV-S	9.303681	10.0	10.0
MLSS-B	6.4036493	0.0	0.0

Taula 5.1: Pesos obtinguts amb l'execució de UEB-1 i UEB-2, aquest últim amb diferent número d'iteracions; tots ells fent servir distancia Euclidea.

Capítol 6

Conclusions

6.1 Algorisme GD

Amb les proves que hem fet concluim que l'algorisme *Gradien Descendant technique* s'ha de refinar més, principalment per dos motius:

- Per raons de cost computacional
- Per resultats obtinguts

Per les execucions hem provat diferents paràmetres. Pels paràmetres η (factor d'aprenentatge) i α hem provats valors dins del rang $[0-1]$ i pel paràmetre *threshold* rangs que oscilen entre 1 i $1x10^{-8}$

Aquest mètode amb els paràmetres assignats o bé no convergeix o ens dona pesos poc satisfactoris.

Cal dir, però, que sempre caldria fer molta més experimentació per poder ajustar el sistema amb els paràmetres adequats.

6.2 Algorismes UEB-1 i UEB-2

6.2.1 Aspectes Generals

Com a primera observació, volem fer notar que tant el UEB-1 com el UEB-2 assigen màxima rellevància (pesos més alts) o mínima rellevància (pesos més baixos) als mateixos atributs. És a dir, mai es dona el fet que un atribut obtingui un pes alt en un algorisme i un pes baix en un altre.

6.2.2 Atributs Rellevants

Als resultats dels algorismes UEB-1 i UEB-2 (veure Taula **??**) podem veure que els atributs rellevants són els nou següents:

- DBO-E, fracció de materia orgànica biodegradable en aigua residual a l'entrada.
- QR-G, caudal de recirculació.
- MCRT-B, edat cel·lular.
- DBO-S, fracció de materia orgànica biodegradable en aigua residual a la sortida.
- SS-E, sòlids en suspensió a l'entrada.
- SSV-E, sòlids volàtils en suspensió a l'entrada.
- SSV-D, sòlids volàtils en suspensió a la decantació.
- SS-S, sòlids en suspensió a la sortida.
- SSV-S, sòlids volàtils en suspensió a la sortida.

6.2.3 Atributs No Rellevants

Per l'algorisme UEB-2, tot el que no és rellevant (pes 10) passa a ser no rellevant (pes 0), el que sembla una mica arriscat. En el cas de UEB-1, és una mica més precis i l'unic atribut que no és rellevant és el següent:

- QA-G, afluència d'aire.

Segons el mateix algorisme (UEB-1), els atributs poc rellevants (amb pes menor a 5) són:

- V30-B, anàlisis volumètric 30.
- PH-E, mesura del pH a la entrada.
- PH-S, mesura del pH a la sortida.

Capítol 7

Línies de Futur

7.1 Futures línies de treball

Dels objectius plantejats ens queden per resoldre dos i que podrien completar perfectament el treball aquí iniciat.

- En primer lloc, i tot i que amb la descripció que teníem de les dades hem procurat fer una interpretació de la rellevància que els algorismes de *Feature Weighting* donaven als atributs en cada cas, seria desitjable que un expert supervises els resultats obtinguts.

- En segon lloc, un estudi molt interessant consistiria a realitzar un procés de clustering a un mateix conjunt de dades fent servir els diferents pesos calculats pels diferents algorismes de *Feature Weighting* i, fins i tot, sense fer servir els pesos.
  Llavors, també seria aconsellable fer servir un mètode de suport a la interpretació de les diferents classificacions obtingudes fent un anàlisi estructural i qualitatiu.

- Sembla un bon punt de partida tenir algun mètode de *feature weighting* implementat, però, com podem obserbar cada mètode dóna diferents resultats, això ens porta a pensar que seria bo tenir algun mètode més implementat per veure quin és el que s'adapta millor al nostre problema.

- L'algorisme *GD* convindria testejar-lo a fons.

- Una vegada verificada la correctesa del *GD* llençar les execucions amb les dades de la depuradora d'aigües.